基因组学数据分析 第四次作业

文献阅读

文章*Diagnosis of multiple cancer types by shrunkencentroids of gene expression*使用数据集为童年时小圆蓝细胞肿瘤（**SRBCT**），其中包含淋巴瘤（Burkitt lymphoma, **BL**）、尤文氏肉瘤（Ewing sarcoma, **EWS**）、成神经细胞瘤（neuroblastoma, **NB**）和横纹肌肉瘤（thabdomyosarcoma, **RMS**）共四种亚型，且每个样本包含`2308`个基因的表达量数据。文章作者将`88`个样本分为含`63`个样本的训练数据集和含`25`个样本的测试数据集。使用类似于*最近质心法(nearest-centroid)*的*最近坍缩质心法(nearest shrunken centroids)*，并筛选了合适的参数，最终达到训练数据与测试数据零错误的成果，并且筛出了`43`个实际参与计算的基因，验证了此方法的效果。

支持向量机

库导入与初始化

```r
library(e1071)
x_train <- read.csv("./homework4-data/xtrain.csv")
y_train <- read.csv("./homework4-data/ytrain.csv")
x_test <- read.csv("./homework4-data/xtest.csv")
y_test <- read.csv("./homework4-data/ytest.csv")
reset_row_names <- function(mat) {
    row.names(mat) <- mat[, 1]
    mat <- mat[, -1]
    return(mat)
}
x_train <- reset_row_names(x_train)
x_test <- reset_row_names(x_test)
y_train <- reset_row_names(y_train)
y_test <- reset_row_names(y_test)
x_train <- as.data.frame(lapply(x_train, as.numeric))
x_test <- as.data.frame(lapply(x_test, as.numeric))
y_train <- as.factor(y_train)
y_test <- as.factor(y_test)
train <- data.frame(x = x_train, y = y_train)
test <- data.frame(x = x_test, y = y_test)
```

> `e1071::svm`函数说明

> `scale`:将数据标准化，使均值为0，方差为1，默认启用
>
> `kernel`:在非线性可分时，引入核函数来做，分为线性核(linear)、多项式核(polynomial)、径向核/高斯核(radial)与sigmoid核(sigmoid)
>
> `cost`:罚分（默认为1），为拉格朗日方法的常数`C`项

`SVM`线性回归

```
svm_fit <- svm(y ~ ., data = train, cost = 10, kernel = "linear", scale =
svm_pred <- predict(svm_fit, newdata = test)
table(svm_fit$fitted, train$y)
```

```
     1  2  3  4
  1  8  0  0  0
  2  0 23  0  0
  3  0  0 12  0
  4  0  0  0 20
```

```
table(svm_pred, test$y)
```

```
svm_pred 1 2 3 4
       1 3 0 0 0
       2 0 6 2 0
       3 0 0 4 0
       4 0 0 0 5
```

可以看到，预测结果中只有两个点预测错误，故粗测总体准确率为90%

参数优化

当然此处数据在上述拟合过程中效果较好，原本即无误分类，故已达到最优点，因此此步分析意义不大。此处仅供流程参考。

```
tuned <- tune(
    svm,
    train.x = x_train,
    train.y = y_train,
    validation.x = x_test,
    validation.y = y_test,
    data = train,
    kernel = "linear",
    scale = FALSE,
```

```r
    ranges = list(
        cost = c(0.1, 0.2, 0.5, 1, 2, 5, 10, 20, 50, 100, 200, 500, 1000)
        gamma = c(0.5, 1, 2, 3, 4)
    )
)
summary(tuned)
```

```

Parameter tuning of 'svm':

- sampling method: 10-fold cross validation

- best parameters:
 cost gamma
  0.1   0.5

- best performance: 0.01428571

- Detailed performance results:
    cost gamma      error dispersion
1  1e-01   0.5 0.01428571  0.0451754
2  2e-01   0.5 0.01428571  0.0451754
3  5e-01   0.5 0.01428571  0.0451754
4  1e+00   0.5 0.01428571  0.0451754
5  2e+00   0.5 0.01428571  0.0451754
6  5e+00   0.5 0.01428571  0.0451754
7  1e+01   0.5 0.01428571  0.0451754
8  2e+01   0.5 0.01428571  0.0451754
9  5e+01   0.5 0.01428571  0.0451754
10 1e+02   0.5 0.01428571  0.0451754
11 2e+02   0.5 0.01428571  0.0451754
12 5e+02   0.5 0.01428571  0.0451754
13 1e+03   0.5 0.01428571  0.0451754
14 1e-01   1.0 0.01428571  0.0451754
15 2e-01   1.0 0.01428571  0.0451754
16 5e-01   1.0 0.01428571  0.0451754
17 1e+00   1.0 0.01428571  0.0451754
18 2e+00   1.0 0.01428571  0.0451754
19 5e+00   1.0 0.01428571  0.0451754
20 1e+01   1.0 0.01428571  0.0451754
21 2e+01   1.0 0.01428571  0.0451754
22 5e+01   1.0 0.01428571  0.0451754
23 1e+02   1.0 0.01428571  0.0451754
24 2e+02   1.0 0.01428571  0.0451754
25 5e+02   1.0 0.01428571  0.0451754
26 1e+03   1.0 0.01428571  0.0451754
27 1e-01   2.0 0.01428571  0.0451754
28 2e-01   2.0 0.01428571  0.0451754
29 5e-01   2.0 0.01428571  0.0451754
30 1e+00   2.0 0.01428571  0.0451754
31 2e+00   2.0 0.01428571  0.0451754
32 5e+00   2.0 0.01428571  0.0451754
33 1e+01   2.0 0.01428571  0.0451754
34 2e+01   2.0 0.01428571  0.0451754
35 5e+01   2.0 0.01428571  0.0451754
36 1e+02   2.0 0.01428571  0.0451754
37 2e+02   2.0 0.01428571  0.0451754
38 5e+02   2.0 0.01428571  0.0451754
```

```
39 1e+03   2.0 0.01428571  0.0451754
40 1e-01   3.0 0.01428571  0.0451754
41 2e-01   3.0 0.01428571  0.0451754
42 5e-01   3.0 0.01428571  0.0451754
43 1e+00   3.0 0.01428571  0.0451754
44 2e+00   3.0 0.01428571  0.0451754
45 5e+00   3.0 0.01428571  0.0451754
46 1e+01   3.0 0.01428571  0.0451754
47 2e+01   3.0 0.01428571  0.0451754
48 5e+01   3.0 0.01428571  0.0451754
49 1e+02   3.0 0.01428571  0.0451754
50 2e+02   3.0 0.01428571  0.0451754
51 5e+02   3.0 0.01428571  0.0451754
52 1e+03   3.0 0.01428571  0.0451754
53 1e-01   4.0 0.01428571  0.0451754
54 2e-01   4.0 0.01428571  0.0451754
55 5e-01   4.0 0.01428571  0.0451754
56 1e+00   4.0 0.01428571  0.0451754
57 2e+00   4.0 0.01428571  0.0451754
58 5e+00   4.0 0.01428571  0.0451754
59 1e+01   4.0 0.01428571  0.0451754
60 2e+01   4.0 0.01428571  0.0451754
61 5e+01   4.0 0.01428571  0.0451754
62 1e+02   4.0 0.01428571  0.0451754
63 2e+02   4.0 0.01428571  0.0451754
64 5e+02   4.0 0.01428571  0.0451754
65 1e+03   4.0 0.01428571  0.0451754
```